myjournal manuscript No.
(will be inserted by the editor)

Small Sample Corrections for LTS and MCD

G. Pison, S. Van Aelst*, and G. Willems

Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium.
e-mail: gpison@uia.ua.ac.be, vaelst@uia.ua.ac.be, gewillem@uia.ua.ac.be

September 10, 2001

Abstract The least trimmed squares estimator and the minimum covariance determinant estimator [5] are frequently used robust estimators of regression and of location and scatter. Consistency factors can be computed for both methods to make the estimators consistent at the normal model. However, for small data sets these factors do not make the estimator unbiased. Based on simulation studies we therefore construct formulas which allow us to compute small sample correction factors for all sample sizes and dimensions without having to carry out any new simulations. We give some examples to illustrate the effect of the correction factor.

Key words Robustness – Least Trimmed Squares estimator – Minimum Covariance Determinant estimator – Bias.

1 Introduction

The classical estimators of regression and multivariate location and scatter can be heavily influenced when outliers are present in the data set. To overcome this problem Rousseeuw [5] introduced the least trimmed squares (LTS) estimator as a robust alternative for least squares regression and the minimum covariance determinant (MCD) estimator instead of the empirical mean and covariance estimators.

Consistency factors can be computed to make the LTS scale and MCD scatter estimators consistent at the normal model. However, these consistency factors are not sufficient to make the LTS scale or MCD scatter unbiased for small sample sizes. Simulations and examples with small sample

* Research Assistant with the FWO Belgium

sizes clearly show that these estimators underestimate the true scatter such that too many observations are identified as outliers.

To solve this problem we construct small sample correction factors which allow us to identify outliers correctly. For several sample sizes n and dimensions p we carried out Monte-Carlo simulations with data generated from the standard Gaussian distribution. Based on the results we then derive a formula which approximates the actual correction factors very well. These formulas allow us to compute the correction factor at any sample size n and dimension p immediately without having to carry out any new simulations.

In Section 2, we focus on the LTS scale estimator. We start with a motivating example and then introduce the Monte-Carlo simulation study. Based on the simulation results we construct the function which yields finite sample corrections for all n and p. Similarly, correction factors for the MCD scatter estimator are constructed in Section 3. The reweighted version of both methods is shortly treated in Section 4. In Section 5 we apply the LTS and MCD on a real data set to illustrate the effect of the small sample correction factor. Section 6 gives the conclusions.

2 Least Trimmed Squares Estimator

Consider the regression model

$$y_i = \boldsymbol{\theta}^t \boldsymbol{x}_i + \varepsilon_i \tag{1}$$

for $i = 1, \ldots, n$. Here $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t \in I\!R^p$ are the regressors, $y_i \in I\!R$ is the response and $\varepsilon_i \in I\!R$ is the error term. We assume that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent of the carriers and are i.i.d. according to $N(0, \sigma^2)$ which is the usual assumption for outlier identification and inference. For every $\boldsymbol{\theta} \in I\!R^p$ we denote the corresponding residuals by $r_i(\boldsymbol{\theta}) = r_i := y_i - \boldsymbol{\theta}^t \boldsymbol{x}_i$ and $r^2_{1:n} \leq \ldots \leq r^2_{n:n}$ denote the squared ordered residuals.

The LTS estimator searches for the optimal subset of size h whose least squares fit has the smallest sum of squared residuals. Formally, for $0.5 \leq \alpha \leq 1$, the LTS estimator $\hat{\boldsymbol{\theta}}$ minimizes the objective function

$$k_\alpha \sqrt{\frac{1}{h(\alpha)} \sum_{i=1}^{h(\alpha)} (r^2)_{i:n}} \tag{2}$$

where $k_\alpha = (\frac{n}{h(\alpha)} \int_{-q}^{q} u^2 d\Phi)^{-1/2}$ with $q = \Phi^{-1}(\frac{\alpha}{2} + \frac{1}{2})$ is the consistency factor for the LTS scale [2] and $h = h(\alpha)$ determines the subset size.

When $\alpha = 0.5$, $h(\alpha)$ equals $[(n + p + 1)/2]$ which yields the highest breakdown value (50%), and when $\alpha = 1$, $h(\alpha)$ equals n, such that we obtain the least squares estimator. For other values of α we compute the subset size by linear interpolation. To compute the LTS we use the FAST-LTS algorithm of Rousseeuw and Van Driessen [8]. The LTS estimate of the error scale is given by the minimum of the objective function (2).

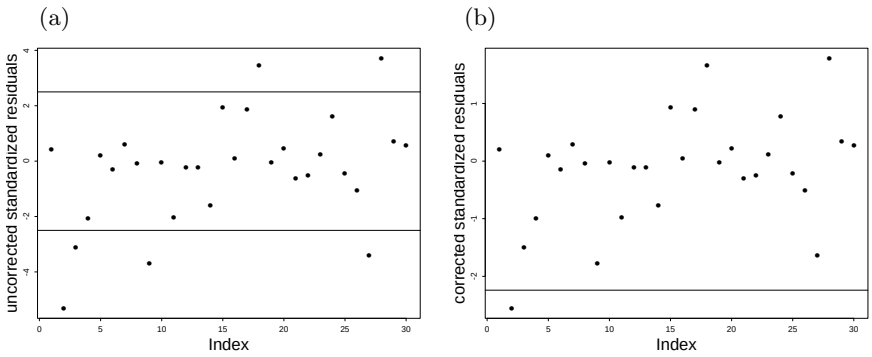

Fig. 1 Robust standardized residuals (a) without correction factors, and (b) with correction factors of a generated data set with $n = 30$ objects and $p = 5$ regressors.

2.1 Example

In this example we generated $n = 30$ points such that the predictor variables are generated from a multivariate standard Gaussian $N_5(0, I)$ distribution and the response variable comes from the univariate standard Gaussian distribution. We used the LTS estimator with $\alpha = 0.5$ to analyse this data set and computed the robust standardized residuals $r_i(\hat{\boldsymbol{\theta}})/\hat{\sigma}$ based on the LTS estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$. Using the cutoff values $+2,5$ and -2.5 we expect to find approximately 1% of outliers in the case of normally distributed errors. Hence, we expect to find at most one outlier in our example. In Figure 1a the robust standardized residuals of the observations are plotted. We see that LTS finds 6 outlying objects which is much more than expected. The main problem is that LTS underestimates the scale of the residuals. Therefore the robust standardized residuals are too large, and too many observations are flagged as outliers.

2.2 Monte-Carlo Simulation Study

To determine correction factors for small data sets, first a Monte-Carlo simulation study is carried out for several sample sizes n and dimensions p. In the simulation we also consider the distribution of $\boldsymbol{x}$ to be Gaussian. Note that the LTS estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^t, \hat{\theta}_2)^t$ with $\hat{\boldsymbol{\theta}}_1$ the slope vector and $\hat{\theta}_2$ the intercept, is regression, scale and affine equivariant (see [6] page 132). This means that

$$\hat{\boldsymbol{\theta}}_1(A\boldsymbol{x}_i, sy_i + \boldsymbol{v}^t\boldsymbol{x}_i + w) = (A^{-1})^t(s\hat{\boldsymbol{\theta}}_1(\boldsymbol{x}_i, y_i) + \boldsymbol{v})$$
$$\hat{\theta}_2(A\boldsymbol{x}_i, sy_i + \boldsymbol{v}^t\boldsymbol{x}_i + w) = s\hat{\theta}_2(\boldsymbol{x}_i, y_i) + w$$

for every $\boldsymbol{v} \in I\!R^p$, $s \neq 0, w \in I\!R$ and nonsingular $p \times p$ matrix A. Also the LTS scale $\hat{\sigma}$ is affine equivariant meaning that

$$\hat{\sigma}^2(sr_i + w) = s^2\hat{\sigma}^2(r_i)$$

for every $s \neq 0$, $w \in I\!R$. From these equivariances it follows that

$$\begin{aligned}&\hat{\sigma}^2(sy_i + \boldsymbol{v}^t\boldsymbol{x}_i + w - \hat{\boldsymbol{\theta}}_1(A\boldsymbol{x}_i, sy_i + \boldsymbol{v}^t\boldsymbol{x}_i + w) - \hat{\theta}_2(A\boldsymbol{x}_i, sy_i + \boldsymbol{v}^t\boldsymbol{x}_i + w)\\ &= s^2\hat{\sigma}^2(y - \hat{\boldsymbol{\theta}}_1(\boldsymbol{x}_i, y_i)\boldsymbol{x}_i - \hat{\theta}_2(\boldsymbol{x}_i, y_i)) \qquad (3)\end{aligned}$$

Therefore it suffices to consider standard Gaussian distributions for $\boldsymbol{x}$ and y since (3) shows that this correction factor remains valid for any Gaussian distribution of $\boldsymbol{x}$ and y.

In the simulation, for sample size n and dimension p we generate regressors $X^{(j)} \in I\!R^{n\times p}$ and a response variable $Y^{(j)} \in I\!R^{n\times 1}$. For each dataset $Z^{(j)} = (X^{(j)}, Y^{(j)})$, $j = 1, \ldots, m$ we then determine the LTS scale $\hat{\sigma}^{(j)}$ of the residuals corresponding to the LTS fit. Finally, the mean $m(\hat{\sigma}) := \frac{1}{m}\sum_{j=1}^{m}\hat{\sigma}^{(j)}$ is computed. If the estimator is unbiased we have $E[\hat{\sigma}] = 1$ for this model, so we expect that also $m(\hat{\sigma})$ equals approximately 1. In general, denote $c^{\alpha}_{p,n} := \frac{1}{m(\hat{\sigma})}$ then $E[c^{\alpha}_{p,n}\hat{\sigma}]$ equals approximately 1, so we can use $c^{\alpha}_{p,n}$ as a finite-sample correction factor to make the LTS scale unbiased.

To determine the correction factor we performed $m = 1000$ simulations for different sample sizes n and dimensions p, and for several values of α. For the model with intercept ($x_p = 1$) we denote the resulting correction factor $c^{\alpha}_{p,n}$ and for the model without intercept it is denoted by $\tilde{c}^{\alpha}_{p,n}$.

From the simulations, we found empirically that for fixed n and p the mean $m(\hat{\sigma})$ is approximately linear in function of α. Therefore we reduced the actual simulations to cases with $\alpha = 0.5$ and $\alpha = 0.875$. For values of α in between we then determine the correction factor by linear interpolation. If $\alpha = 1$ then least squares regression is carried out. In this case, we don't need a correction factor because this estimator is unbiased. So, if $0.875 \leq \alpha \leq 1$ we interpolate between the value of $m(\hat{\sigma})$ for $\alpha = 0.875$ and 1 to determine the correction factor.

In Table 1, the mean $m(\hat{\sigma})$ for LTS with intercept and $\alpha = 0.5$ is given for several values of n and p. We clearly see that when the sample size n is small, $m(\hat{\sigma})$ is very small. Moreover, for n fixed, the mean becomes smaller when the dimension increases. Note that for fixed p the mean increases monotone to 1, so for large samples the consistency factor suffices to make the estimator unbiased. Table 2 shows the result for $\alpha = 0.875$. In comparison with Table 1 we see that these values of $m(\hat{\sigma})$ are higher such that the correction factor $c^{0.875}_{p,n}$ will be smaller than $c^{0.5}_{p,n}$ for the same value of n and p. Similar results were found for LTS without intercept.

2.3 Finite Sample Corrections

We now construct a function which approximates the actual correction factors obtained from the simulations and allows us to compute the correction

Table 1 $m(\hat{\sigma})$ for $\alpha = 0.5$ and for several sample sizes n and dimensions p

p \ n	20	25	30	35	50	55	80	85	100
1	0.71	0.77	0.77	0.81	0.84	0.86	0.89	0.90	0.91
3	0.49	0.58	0.60	0.65	0.71	0.74	0.79	0.81	0.83
5	0.35	0.45	0.46	0.53	0.60	0.64	0.71	0.72	0.75
8	0.25	0.26	0.34	0.36	0.49	0.51	0.62	0.62	0.67

Table 2 $m(\hat{\sigma})$ for $\alpha = 0.875$ and for several sample sizes n and dimensions p

p \ n	20	25	30	35	50	55	80	85	100
1	0.91	0.94	0.94	0.96	0.97	0.97	0.98	0.98	0.99
3	0.83	0.86	0.88	0.90	0.93	0.94	0.95	0.96	0.97
5	0.73	0.77	0.83	0.83	0.89	0.90	0.93	0.93	0.95
8	0.56	0.69	0.72	0.75	0.84	0.85	0.90	0.90	0.92

factor at any sample size n and dimension p immediately without having to carry out any new simulations. First, for a fixed dimension p we plotted the mean $m(\hat{\sigma})$ versus the number of observations n. We made plots for several dimensions p ($1 \leq p \leq 10$), for $\alpha = 0.5$, and $\alpha = 0.875$ and for LTS with and without intercept. Some plots are shown in Figure 2.

From these plots we see that for p fixed the mean $m(\hat{\sigma})$ has a smooth pattern in function of n. For fixed p we used the model

$$f_p^\alpha(n) = 1 + \frac{\gamma}{n^\beta} \tag{4}$$

to fit the mean $m(\hat{\sigma})$ in function of n. Hence, for each p and α we obtain the corresponding parameters $\gamma := \gamma_{p,\alpha}$ and $\beta := \beta_{p,\alpha}$ for LTS with intercept and $\gamma := \tilde{\gamma}_{p,\alpha}$, $\beta := \tilde{\beta}_{p,\alpha}$ for LTS without intercept. In Figure 2 the functions obtained by using the model (4) are superimposed. We see that the function values $f_p^\alpha(n)$ approximate the actual values of $m(\hat{\sigma})$ obtained from the simulations very well.

When the regression dataset has a dimension that was included in our simulation study, then the functions $f_p^\alpha(n)$ already yield a correction factor for all possible values of n. However, when the data set has another dimension, then we have not yet determined the corresponding correction factor. To be able to obtain correction factors for these higher dimensions we fitted the function values $f_p^\alpha(qp^2)$ for $q = 3$ and $q = 5$ as a function of the number of dimensions p ($p \geq 2$). In Figure 3 we plotted the values $f_p^\alpha(qp^2)$ versus the dimension p for the LTS with intercept and $\alpha = 0.5$. Also in Figure 3 we see a smooth pattern. Note that the function values $f_p^\alpha(qp^2)$ converge to 1 as p goes to infinity since we know from (4) that $f_p^\alpha(qp^2)$ goes to 1 if qp^2 goes to infinity. The model we used to fit the values $f_p^\alpha(qp^2)$ in function of

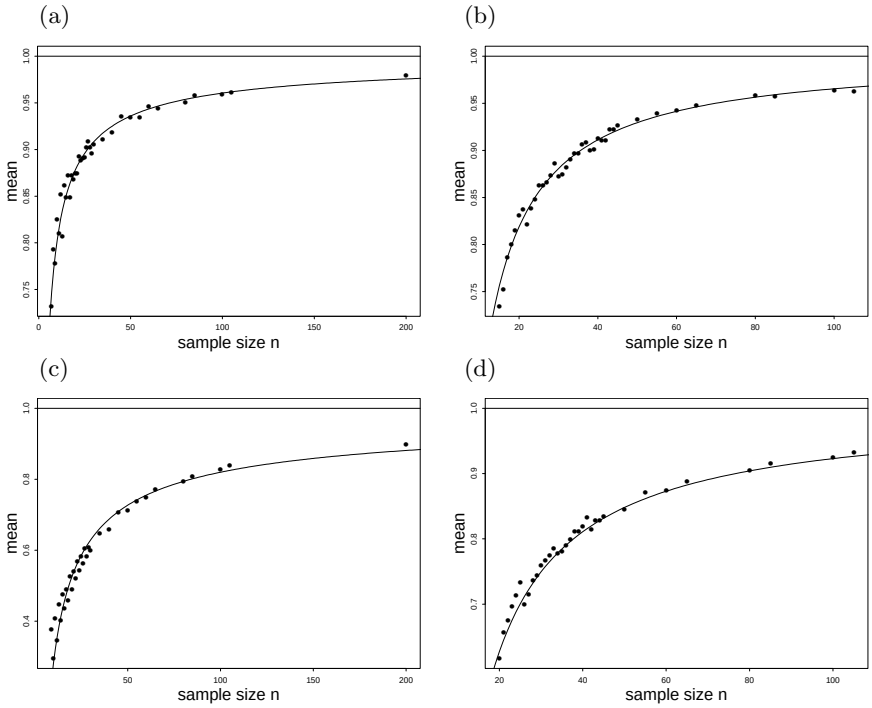

Fig. 2 The approximating function $f_p^\alpha(n)$ for (a) $p = 1$, $\alpha = 0.5$ and LTS without intercept, (b) $p = 4$, $\alpha = 0.875$ and LTS without intercept , (c) $p = 3$, $\alpha = 0.5$ and LTS with intercept, (d) $p = 7$, $\alpha = 0.875$ and LTS with intercept.

p is given by

$$g_q^\alpha(p) = 1 + \frac{\eta}{p^\kappa}. \tag{5}$$

By fitting this model for $q = 3$ and 5 and $\alpha = 0.5$ and 0.875 we obtain the corresponding parameters $\eta := \eta_{q,\alpha}$ and $\kappa := \kappa_{q,\alpha}$ for LTS with intercept and $\eta := \tilde{\eta}_{q,\alpha}$, $\kappa := \tilde{\kappa}_{q,\alpha}$ for LTS without intercept. From Figure 3 we see that the resulting functions fit the points very well.

Finally, for any n and p we now have the following procedure to determine the corresponding correction factor for the LTS scale estimator. For the LTS with intercept the correction factor in the case $p = 1$ is given by $c_{1,n}^\alpha := \frac{1}{f_1^\alpha(n)}$ where $f_1^\alpha(n) = 1 + \gamma_{1,\alpha}/n^{\beta_{1,\alpha}}$. In the case $p > 1$, we first solve the following system of equations

$$1 + \frac{\eta_{3,\alpha}}{p^{\kappa_{3,\alpha}}} = 1 + \frac{\gamma_{p,\alpha}}{(3p^2)^{\beta_{p,\alpha}}} \tag{6}$$

$$1 + \frac{\eta_{5,\alpha}}{p^{\kappa_{5,\alpha}}} = 1 + \frac{\gamma_{p,\alpha}}{(5p^2)^{\beta_{p,\alpha}}} \tag{7}$$

to obtain the estimates $\hat{\gamma}_{p,\alpha}$ and $\hat{\beta}_{p,\alpha}$ of the parameter values $\gamma_{p,\alpha}$ and $\beta_{p,\alpha}$. Note that the system of equations (6)–(7) can be rewritten into a linear sys-

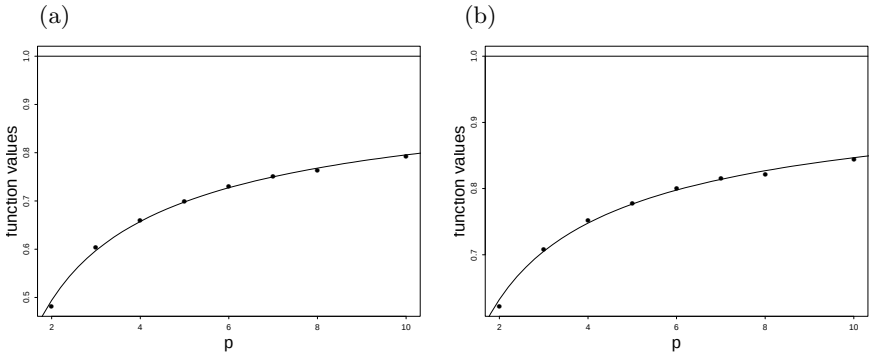

Fig. 3 The approximating function $g_q^\alpha(p)$ for (a) $q = 3$, $\alpha = 0.5$ and LTS with intercept, (b) $q = 5$, $\alpha = 0.5$ and LTS with intercept.

tem of equations by taking logarithms. The corresponding correction factor is then given by $c_{p,n}^\alpha := 1/\hat{f}_p^\alpha(n)$ where $\hat{f}_p^\alpha(n) = 1 + \hat{\gamma}_{p,\alpha}/n^{\hat{\beta}_{p,\alpha}}$. Similarly, we also obtain the correction factors for the LTS without intercept.

Using this procedure we obtain the functions shown in Figure 4. We can clearly see that these functions are nearly the same as the original functions $f_p^\alpha(n)$ shown in Figure 2.

Let us reconsider the example of Section 2.1. The corrected LTS estimator with $\alpha = 0.5$ is now used to analyse the dataset. The resulting robust standardized residuals are plotted in Figure 1b. Using the cutoff values $\Phi^{-1}(0.9875)$ and $-\Phi^{-1}(0.9875)$ we find 1 outlier which corresponds with the 2.5% of outliers we expect to find. Also, we clearly see that the corrected residuals are much smaller than the uncorrected. The corrected residuals range between -3 and 2 while the uncorrected residuals range between -5 and 4. We conclude that the scale is not underestimated when we use the LTS estimator with small sample corrections and therefore it gives more reliable values for the standardized residuals and more reliable outlier identification.

Finally, we investigated whether the correction factor is also valid when working with non-normal explanatory variables. In Table 3 we give the mean $m(\hat{\sigma})$ for some simulation set ups where we used exponential, student (with 3 df.) and cauchy distributed carriers. The approximated values $\hat{f}_p^\alpha(n)$ of $m(\hat{\sigma})$ obtained with normally distributed carriers are given between brackets. From Table 3 we see that the difference between the simulated value and the correction factor is very small. Therefore, we conclude that in general, also for nonnormal carrier distributions, the correction factor makes the LTS scale unbiased.

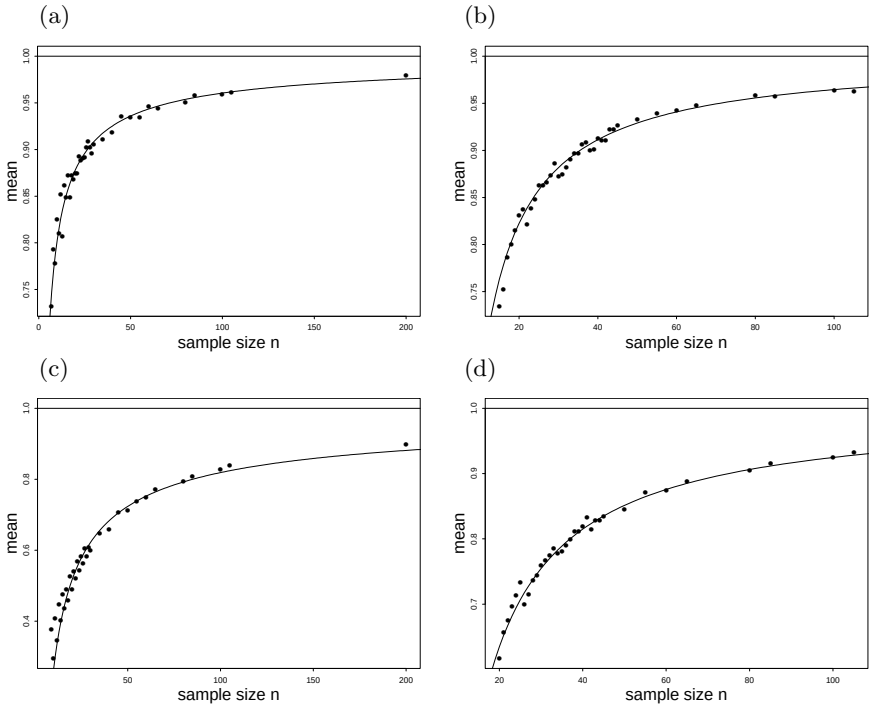

Fig. 4 The approximation $\hat{f}_p^{\alpha}(n)$ for (a) $p = 1$, $\alpha = 0.5$ and LTS without intercept, (b) $p = 4$, $\alpha = 0.875$ and LTS without intercept , (c) $p = 3$, $\alpha = 0.5$ and LTS with intercept, (d) $p = 7$, $\alpha = 0.875$ and LTS with intercept.

Table 3 $m(\hat{\sigma})$ for several other distributions of the carriers.

	$n = 20$	$n = 40$	$n = 60$	$n = 80$
exp, $p = 4$, $\alpha = 0.875$, without intercept	0.84 (0.82)	0.91 (0.91)	0.94 (0.94)	0.96 (0.96)
t_3, $p = 3$, $\alpha = 0.5$, with intercept	0.50 (0.52)	0.67 (0.68)	0.75 (0.75)	0.80 (0.79)
cauchy, $p = 7$, $\alpha = 0.875$, with intercept	0.63 (0.63)	0.83 (0.81)	0.88 (0.88)	0.92 (0.91)

3 Minimum Covariance Determinant Estimator

The MCD estimates the location vector $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$. Suppose we have a dataset $Z_n = \{\boldsymbol{z}_i; i = 1, \ldots, n\} \subset \mathbb{R}^p$, then the MCD searches for the subset of $h = h(\alpha)$ observations whose covariance matrix has the lowest determinant. For $0.5 \leq \alpha \leq 1$, its objective is to minimize the determinant of

$$l_{\alpha} S_{full} \tag{8}$$

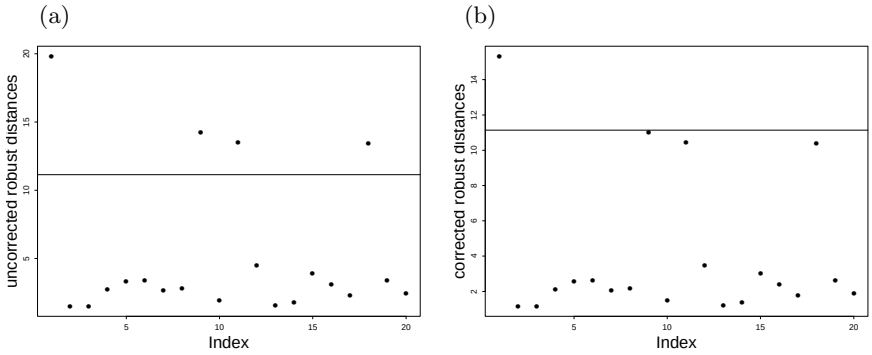

Fig. 5 Robust distances (a) without correction factors, (b) with correction factors, of a generated data set with $n = 20$ objects and $p = 4$ dimensions.

where $S_{full} = \frac{1}{h(\alpha)} \sum_{i=1}^{h(\alpha)} (z_i - \hat{\mu}_n)(z_i - \hat{\mu}_n)^t$ with $\hat{\mu}_n = \frac{1}{h(\alpha)} \sum_{i=1}^{h(\alpha)} z_i$. The factor $l_\alpha = \alpha / F_{\chi^2_{p+2}}(q_\alpha)$ with $q_\alpha = \chi^2_{p,\alpha}$ makes the MCD scatter estimator consistent at the normal model (see [3]). The MCD center is then the mean of the optimal subset and the MCD scatter is a multiple of its covariance matrix as given by (8). A fast algorithm have been constructed to compute the MCD ([7])

3.1 Example

Similarly as for LTS, we generated data from a multivariate standard Gaussian distribution. For $n = 20$ observations of $N_4(0, I)$ we computed the MCD estimates with $\alpha = 0.75$. As cutoff value to determine outliers the 97.5% quantile of the χ^2_4 distribution is used. Since no outliers are present, we therefore expect that MCD will find at most one outlier in this case. Nevertheless, the MCD estimator identifies 4 outlying objects as shown in Figure 5a where we plotted the robust distances of the 20 observations. Hence a similar problem arises as with LTS. The MCD estimator underestimates the volume of the scatter matrix, such that the robust distances are too large. Therefore the MCD identifies too many observations as outliers.

3.2 Monte-Carlo Simulation Study

A Monte-Carlo simulation study is carried out for several sample sizes n and dimensions p. We generated datasets $\boldsymbol{X}^{(j)} \in I\!R^{n \times p}$ from the standard Gaussian distribution. It suffices to consider the standard Gaussian distribution since the MCD is affine equivariant (see [6] page 262). For each dataset $\boldsymbol{X}^{(j)}$, $j = 1, \ldots, m$ we then determine the MCD scatter matrix $\hat{\boldsymbol{\Sigma}}^{(j)}$. If the estimator is unbiased, we have that $E[\hat{\boldsymbol{\Sigma}}] = I_p$ so we expect that the p-th root of the determinant of $\hat{\Sigma}$ equals 1. Therefore, the mean of the p-th root of

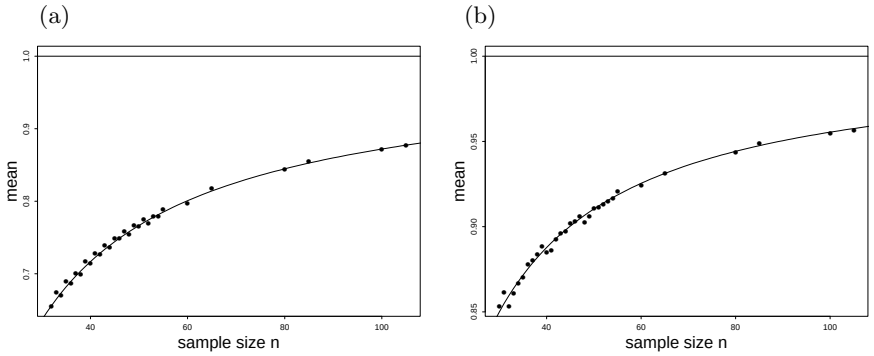

Fig. 6 The approximation $\hat{f}_p^{\alpha}(n)$ for (a) $p = 8$, $\alpha = 0.5$, and (b) $p = 6$, $\alpha = 0.875$.

the determinant given by $m(|\hat{\boldsymbol{\Sigma}}|) := \frac{1}{m}\sum_{j=1}^{m}(|\hat{\boldsymbol{\Sigma}}^{(j)}|)^{1/p}$, where $|A|$ denotes the determinant of a square matrix A, is computed. Denote $d_{p,n}^{\alpha} := \frac{1}{m(|\hat{\boldsymbol{\Sigma}}|)}$, then we expect that the determinant of $d_{p,n}^{\alpha}\hat{\boldsymbol{\Sigma}}$ equals approximately 1. Similarly as for LTS, we now use $d_{p,n}^{\alpha}$ as a finite-sample correction factor for MCD. We performed $m = 1000$ simulations for different sample sizes n and dimensions p, and for several values of α to compute the correction factors.

From the simulation study similar results as for LTS were obtained. Empirically we found that the mean $m(|\hat{\boldsymbol{\Sigma}}|)$ is approximately linear in function of α so we reduced the actual simulations to cases with $\alpha = 0.5$ and $\alpha = 0.875$. The other values of α are determined by linear interpolation. Also here we saw that the mean is very small when the sample size n is small, and for fixed p the mean increases monotone to 1 when n goes to infinity.

3.3 Finite Sample Corrections

We now construct a function which approximates the actual correction factors obtained from the simulations. The same setup as for LTS is used. Model (4) and for $p > 2$ also model (5) with $q = 2$ and $q = 3$ are used to derive a function which yields a correction factor for every n and p. The function values $\hat{f}_p^{\alpha}(n)$ obtained from this procedure are illustrated in Figure 6. In this Figure the mean $m(|\hat{\boldsymbol{\Sigma}}|)$ is plotted versus the sample size n for a fixed p and α and superimposed are the functions $\hat{f}_p^{\alpha}(n)$. We see that the function values $\hat{f}_p^{\alpha}(n)$ are very close to the original values obtained from the simulations.

Finally, we return to the example in Section 3.1. We now use the corrected MCD estimator to analyse the dataset. The resulting robust distances are plotted in Figure 5b. Using the same cutoff value we now find 1 outlier which corresponds to the 2.5% of outliers that is expected. Note that the corrected distances are much smaller than the uncorrected ones. The

corrected distances are all below 15.5 while the uncorrected distances range between 0 and 20. When we use the MCD with small sample corrections the volume of the MCD scatter estimator is not underestimated anymore, so we obtain more reliable robust distances and outlier identification.

4 Reweighted LTS and MCD

To increase the efficiency of the LTS and MCD, the reweighted version of these estimators is often used in practice [6]. Similarly to the initial LTS and MCD, the reweighted LTS scale and MCD scatter are not unbiased at small samples even when the consistency factor is included. Therefore, we also determine small sample corrections for the reweighted LTS and MCD based on the corrected LTS and MCD as initial estimators. We performed Monte-Carlo studies similar to those for the initial LTS and MCD to compute the finite-sample correction factor for several sample sizes n and dimensions p. Based on these simulation results, we then constructed functions which determine the finite sample correction factor for all n and p.

5 Examples

Let us now look at some real data examples. First we consider the Coleman data set which contains information on 20 schools from the Mid-Atlantic and New England states, drawn from a population studied by [1]. The dataset contains 5 predictor variables which are the staff salaries per pupil (x_1), the percent of white-collar fathers (x_2), the socioeconomic status composite deviation (x_3), the mean teacher's verbal test score (x_4) and the mean mother's educational level (x_5). The response variable y measures the verbal mean test score. Analyzing this dataset using LTS with intercept and $\alpha = 0.5$, we obtain the standardized residuals shown in Figure 7. Figure 7a is based on LTS without correction factor while Figure 7b is based on the corrected LTS. The corresponding results for the reweighted LTS are shown in Figures 7c and 7d. Based on the uncorrected LTS 7 objects are identified as outliers. On the other hand, by using the corrected LTS the standardized residuals are rescaled and only 2 huge outliers and 1 boundary case are left. The standardized residuals of the uncorrected LTS range between -11 and 15 while the values of the corrected LTS range between -4 and 5. Also when using the reweighted LTS we can see that the uncorrected LTS finds 5 outliers and 2 boundary cases while the corrected version only finds 2 outliers.

In the second example we consider the aircraft dataset [4] which deals with 23 single-engine aircraft built between 1947–1979. We use the MCD with $\alpha = 0.75$ to analyse the 4 independent variables which are Aspect Ratio (x_1), Lift-to-Drag ratio (x_2), Weight (x_3) and Thrust (x_4). Based on MCD without correction factor we obtain the robust distances shown in Figure 8a. We see that 4 observations are identified as outliers of which

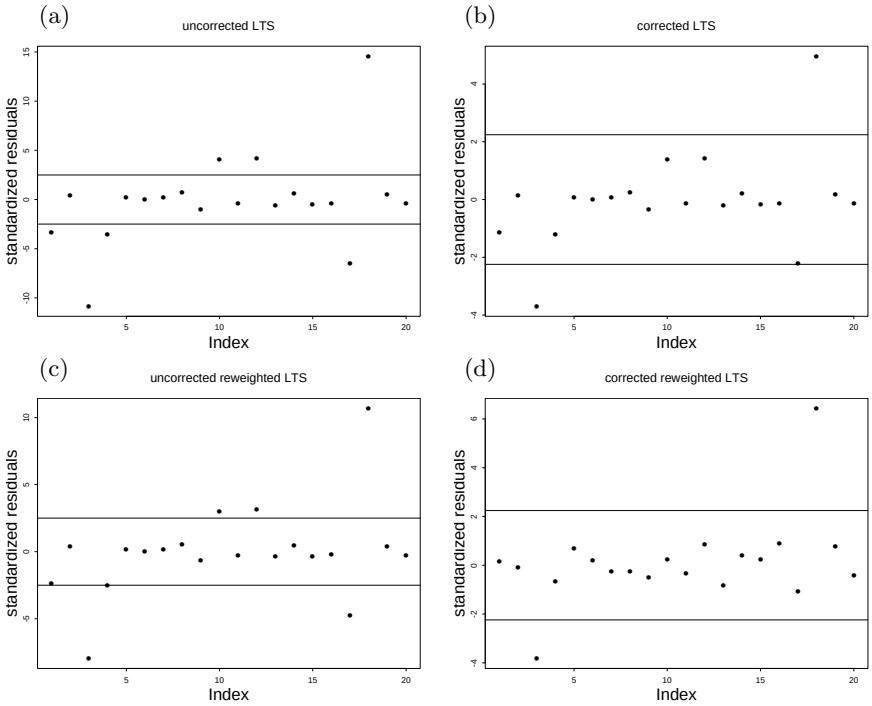

Fig. 7 Robust standardized residuals for the coleman data ($n = 20$, $p = 5$) based on LTS with intercept and $\alpha = 0.5$ (a) uncorrected , (b) corrected, (c) uncorrected reweighted, and (d) corrected reweighted .

aircraft 15 is a boundary case. The robust distance of aircraft 14 equals 494. If we use the corrected MCD then we obtain the robust distances in Figure 8b where the boundary case has disappeared. Note that the robust distances have been rescaled. For example the robust distance of aircraft 14 is reduced to 395. Similar results are obtained for the reweighted MCD as shown by Figures 8c and 8d.

6 Conclusions

Even when a consistency factor is included, this is not sufficient to make the LTS and MCD unbiased at small samples. Consequently, the LTS based standardized residuals and the MCD based robust distances are too large such that too many observations are identified as outliers. To solve this problem, we performed Monte-Carlo simulations to compute correction factors for several sample sizes n and dimensions p. Based on the simulation results we constructed functions that allow us to determine the correction factor for all sample sizes and all dimensions. Similar results have been obtained for the reweighted LTS and MCD. Some examples have been given to illustrate the difference between the uncorrected and corrected estimators.

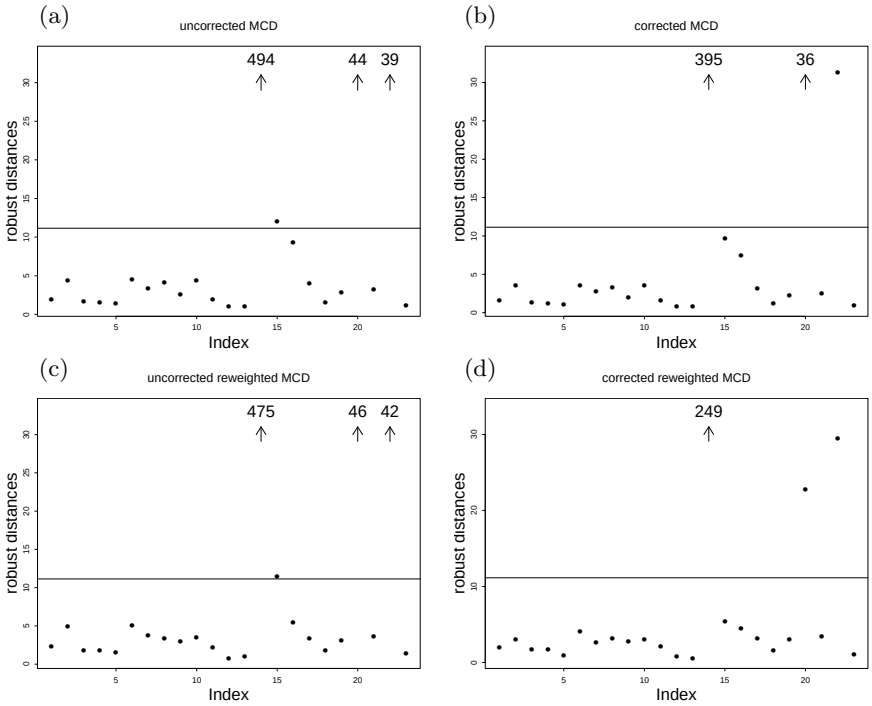

Fig. 8 Robust distances for the aircraft data ($n = 23$, $p = 4$) based on MCD with $\alpha = 0.75$ (a) uncorrected , (b) corrected, (c) uncorrected reweighted, and (d) corrected reweighted.

References

1. J. Coleman et al., *Equality of Educational Opportunity* (U.S. Department of Health, Washington D.C. 1966).
2. C. Croux and P.J. Rousseeuw, "A Class of High-Breakdown Scale Estimators Based on Subranges," Communications in Statistics **21,** (1992), 1935–1951.
3. C. Croux and G. Haesbroeck, "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," Journal of Multivariate Analysis **71,** (1999) 161–190.
4. J.B. Gray, "Graphics for Regression Diagnostics," American Statistical Association Proceedings of the Statistical Computing Section, (1985) 102–107.
5. P.J. Rousseeuw, "Least Median of Squares Regression," Journal of the American Statistical Association **79,** (1984) 871–880.
6. P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection* (Wiley-Interscience, New York 1987).
7. P.J. Rousseeuw and K. Van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics **41,** (1999) 212–223.
8. P.J. Rousseeuw and K. Van Driessen, "Computing LTS Regression for Large Data sets," Technical Report (1999), Universitaire Instelling Antwerpen.

.